

## URBAN FINDINGS

# Validity of Food Outlet Databases from Commercial and Community Science datasets in Vancouver and Montreal

Caislin Firth<sup>1</sup> , Geneva Beairisto<sup>2</sup>, Colin Ferster<sup>2</sup>, Grace Longson<sup>3</sup>, Kevin Manaugh<sup>3</sup>, Yan Kestens<sup>4</sup>, Meghan Winters<sup>2</sup><sup>1</sup> Psychiatry & Behavioral Sciences, University of Washington, <sup>2</sup> Faculty of Health Sciences, Simon Fraser University, <sup>3</sup> Department of Geography & Bieler School of Environment, McGill University, <sup>4</sup> Centre de recherche du CHUM, Université de Montréal

Keywords: Food environment, community science, citizen science, food retail, geocoding, validation study, urban health, foodscapes

<https://doi.org/10.32866/001c.35619>

---

## Findings

---

We conducted a case study to assess the validity of community science (Yelp, OpenStreetMaps) and commercial (DMTI) food outlet datasets. We compared counts of food outlets from 13 street segments in Vancouver and Montreal to Google Street View. We found that OpenStreetMaps correctly identified the most outlets in both cities and DMTI consistency overcounted outlets. In Vancouver, we assessed validity by outlet type, again OpenStreetMap performed the best overall but largely missed grocery stores, and Yelp did not include convenience stores. Results provide insights into using different commercial and open-source datasets to measure food environments.

## 1. QUESTIONS

Studies of neighbourhood food environments typically rely on commercial or registry-based data systems. Previous work has focused on data quality and geographical biases in commercial data sources (Lebel et al. 2017; Daepp and Black 2017; Clary and Kestens 2013). Yet it is unclear whether community science food outlet data are a reliable alternative to costly commercial datasets, which can be used by academics, practitioners, and policy makers to understand food environments in real time. To understand the utility of commercial and community science data, overall and for counts of specific food outlet categories, we calculated the validity of food outlet data from commercial data (DMTI Enhanced Points of Interest) and community science data (OpenStreetMap (OSM) and Yelp) via comparisons to Google Street View (GSV), the ‘reference standard’.

## 2. METHODS

We conducted a case study in two Canadian cities: Vancouver and Montreal; using 2020 DMTI Enhanced Points of Interest dataset (Supplemental Table 1) and obtained Yelp data (Supplemental Table 2) and OSM data (Supplemental Table 3) in Summer 2021. The DMTI Enhanced Points of Interest is a proprietary dataset commonly used for research, though its validity varies substantially, and food outlets in DMTI were modestly correlated with food outlets from Canadian tax records (Stevenson et al. 2022). Both OSM and Yelp are community science datasets. OSM uses crowdsourced data to populate maps that are free to use, which has attracted attention from research communities (“OpenStreetMap Research” 2022). Yelp is a public company that publishes crowd-sourced business reviews. Yelp data can be used for research purposes, but data access is restricted.

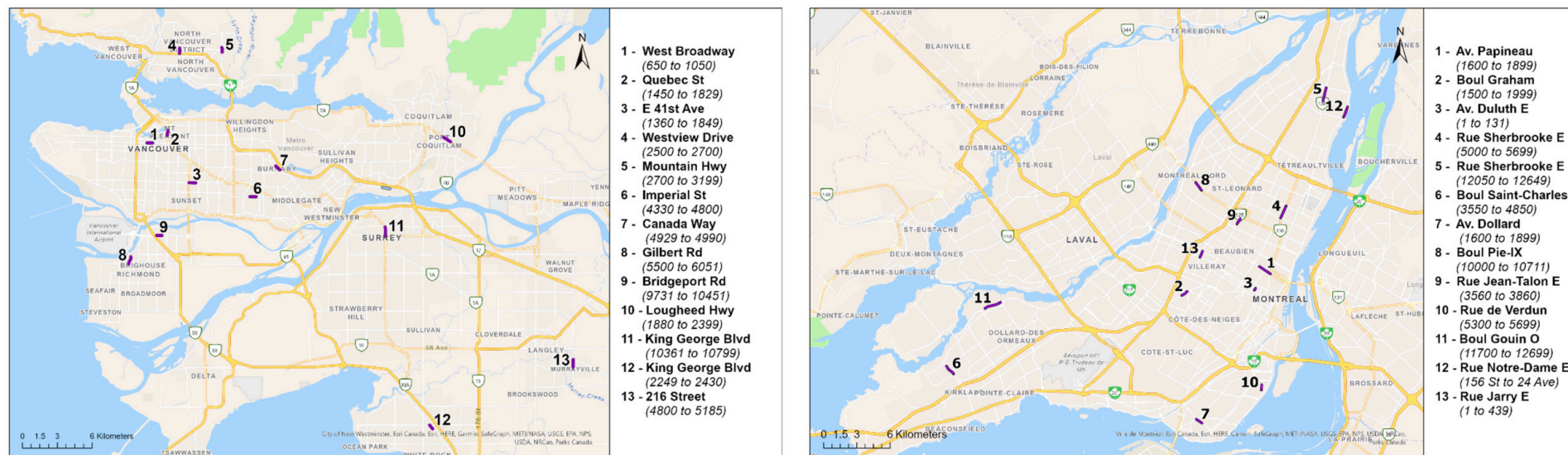


Figure 1. Locations of street segments in Vancouver Metro area and Montreal

*Note: Each street segment is between 200 to 1,000 meters in length.*

We compared the number of outlets in DMTI, OSM, and Yelp to GSV for a sample of 13 randomly selected street segments each across the Island of Montreal and the Vancouver Metropolitan Area ([Figure 1](#)). Eligible street segments were 200 to 1,000 meters long with at least two food outlets identified in GSV. Sampled segments with  $\leq 1$  food outlets were resampled from a generated list of street segments in each city. Along each street segment, we recorded the name and location of food outlets in each dataset. We identified food outlets as buildings used primarily for the purpose of selling food and beverages for on and off-premise consumption. We considered seven categories of outlets (grocery, convenience, cafes, bakeries/dessert shops, bars/liquor stores, restaurants, fast food) and included outlets that sold alcoholic beverages because of public health harms associated with access to alcohol outlets (Bright et al. 2018). For the comparisons, GSV was used as the reference standard, as previous work has shown it accurately identifies street-level environmental features (Steinmetz-Wood et al. 2019) and food outlets when compared to in-person fieldwork (de Menezes et al. 2020). We considered a food outlet to be correctly identified if it had the same or similar name, location, and was an operating business on GSV. Analysis completed September 2021.

To evaluate the validity of each dataset, we calculated two validity measures and their corresponding 95% confidence intervals using Clopper-Pearson exact method for binomial probability: 1) Sensitivity (*i.e., the percentage of actual food outlets that are present in the dataset*), and 2) Positive predictive value (PPV) (*i.e., the percentage of food outlets in the dataset that are actually food outlets*). Validity measures were calculated for each city and by food outlet categories in Vancouver (grocery, convenience, café, bakeries/dessert shops, bars/liquor stores, restaurants, fast food, supplemental Tables 1-3 for detail). It is important to consider multiple measures of validity because sensitivity alone does not account for false negative results. For example, low sensitivity and high PPV means that the dataset did not identify all food outlets but for the outlets in the dataset, most were correct.

### 3. FINDINGS

The number of food outlets per segment in Vancouver was 15 outlets on average (range 2-38), and 10 in Montreal (range 2-20). In both cities, OSM and Yelp undercounted the total number of food outlets and DMTI overcounted the number of food outlets in Vancouver by 32 and undercounted in Montreal by 14 ([Table 1](#)).

**Validity of food outlet datasets.** In Vancouver, OSM had the highest sensitivity and PPV, meaning that most outlets identified in OSM were present in GSV and OSM captured the largest proportion of food outlets that were in GSV ([Table 1](#)). Of the 194 food outlets in Vancouver, OSM correctly identified 121 (sensitivity 62%, 95% CI: 56%,70%). In Montreal, there were 133 food outlets and OSM correctly identified 71—which corresponded with a lower

Table 1. DMTI, OSM, and Yelp food outlet data validity for a sample of street segments in Vancouver Metro area and Montreal

|                  | Total food outlets in dataset        | Food outlets present in dataset and GSV | Food outlets in GSV and not in dataset | Food outlets present in dataset absent in GSV | % of food outlets in GSV present in dataset | % of food outlets in dataset present in GSV          |
|------------------|--------------------------------------|-----------------------------------------|----------------------------------------|-----------------------------------------------|---------------------------------------------|------------------------------------------------------|
|                  | <i>Total food outlets in dataset</i> | <i>True positives (TP)</i>              | <i>False Negatives (FN)</i>            | <i>False positives (FP)</i>                   | <i>Sensitivity (TP/TP+FN) (95% CI)</i>      | <i>Positive Predictive Value (TP/TP+FP) (95% CI)</i> |
| <b>Vancouver</b> |                                      |                                         |                                        |                                               |                                             |                                                      |
| GSV              | 192                                  | 192                                     | 0                                      | 0                                             | 100% (98%, 100%)                            | 100% (98%, 100%)                                     |
| DMTI             | 223                                  | 79                                      | 113                                    | 144                                           | 41% (34%, 48%)                              | 35% (29%, 42%)                                       |
| OSM              | 136                                  | 121                                     | 71                                     | 15                                            | 62% (56%, 70%)                              | 89% (82%, 94%)                                       |
| Yelp             | 106                                  | 86                                      | 106                                    | 20                                            | 45% (38%, 52%)                              | 81% (72%, 88%)                                       |
| <b>Montreal</b>  |                                      |                                         |                                        |                                               |                                             |                                                      |
| GSV              | 133                                  | 133                                     | 0                                      | 0                                             | 100% (97%, 100%)                            | 100% (97%, 100%)                                     |
| DMTI             | 119                                  | 92                                      | 41                                     | 27                                            | 69% (61%, 77%)                              | 77% (69%, 84%)                                       |
| OSM              | 80                                   | 71                                      | 62                                     | 9                                             | 53% (45%, 62%)                              | 89% (80%, 95%)                                       |
| Yelp             | 113                                  | 96                                      | 37                                     | 17                                            | 72% (64%, 80%)                              | 85% (77%, 91%)                                       |

sensitivity (53%, 95% CI: 45%, 62%) compared to the other datasets, and a higher PPV (89%, 95% CI: 80%, 95%). Validity measures varied substantially across street segments ([Figure 2](#)). OSM had accurate counts of all food outlets for 10 of the 26 street segments, Yelp had eight, and DMTI three—all in Montreal.

Both DMTI and Yelp performed better in Montreal relative to Vancouver. In Montreal, both DMTI and Yelp identified a similar number of stores with relatively few false positives. In Vancouver, Yelp identified fewer outlets than GSV; whereas DMTI overcounted, with two-thirds of stores in DMTI not identified in GSV. The large number of ‘extra’ food outlets may include outlets that closed between DMTI data release (2020) and the ascertainment of other datasets (Summer 2021). Data inaccuracies (e.g. different addresses between datasets, duplicate entries) were more common in DMTI. Examining data inaccuracies at the record level is time consuming and may be not feasible for larger scale projects or without local knowledge. In addition, DMTI included non-food businesses (e.g. daycares) despite selecting businesses with food-related Standardized Industry Codes (Supplemental Table 3).

**Validity by food outlet type** Some research requires data on specific food outlets, thus we explored dataset validity across seven categories of food outlets in Vancouver. OSM correctly identified the most outlets across types, though

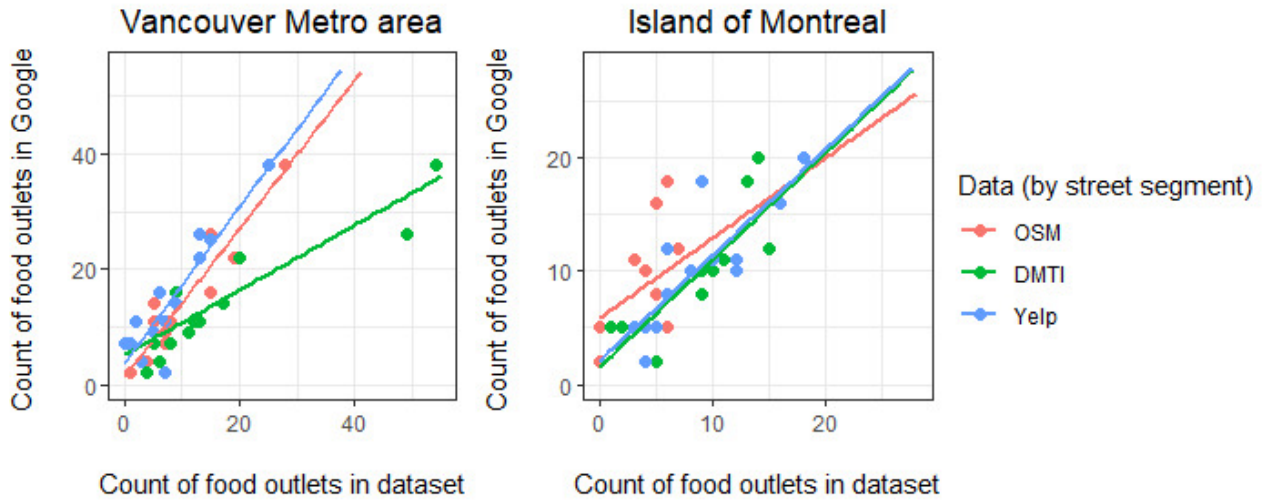


Figure 2. Food outlets in DMTI, OSM and Yelp compared to Google Street View for street segments in Vancouver Metro area and Montreal

Table 2. DMTI, OSM, and Yelp food outlet data validity by outlet type for a sample of street segments in Vancouver Metro area

| Food Outlet Type       | GSV (Count) | DMTI (Count) | OSM (Count) | Yelp (Count) | DMTI PPV | OSM PPV | Yelp PPV |
|------------------------|-------------|--------------|-------------|--------------|----------|---------|----------|
| Grocery                | 25          | 41           | 8           | 15           | 22%      | 88%     | 67%      |
| Convenience            | 16          | 19           | 20          | 1            | 68%      | 75%     | 0%       |
| Cafes                  | 28          | 18           | 25          | 8            | 44%      | 83%     | 75%      |
| Bakeries/Dessert shops | 9           | 10           | 5           | 4            | 40%      | 80%     | 75%      |
| Bars/Liquor Stores     | 13          | 9            | 12          | 4            | 67%      | 64%     | 25%      |
| Restaurants            | 77          | 126          | 46          | 61           | 17%      | 85%     | 79%      |
| Fast food              | 24          | -            | 22          | 13           | -        | 77%     | 54%      |

*Note: The DMTI dataset did not include fast food outlets. The restaurants food type for DMTI therefore includes both full-service restaurants and limited-service eating places (i.e., fast food businesses).*

OSM missed most grocery stores ([Table 2](#)). DMTI suffered from overcounting outlets and only identified 17% of restaurants on GSV. Yelp consistently undercounted across outlet types and, importantly, did not identify any convenience stores. These results provide insights into how to pair research questions with the most appropriate dataset.

Our analysis sheds lights on the validity of business registries – both commercial (DMTI) and open data (OSM, Yelp) – at counting and categorizing food outlets for a sample of streetscapes in two Canadian cities. The results can inform how food environment practitioners and researchers can use data registries in their work. We encourage future users of community science datasets to incorporate validity assessments, to characterize how



availability of data and validity measures vary across space, as previous work found that Yelp data coverage is clustered within specific neighbourhoods (Folch, Spielman, and Manduca 2018). While we only assessed 13 street segments in each city, these segments were in diverse neighbourhood environments. We acknowledge that our matching criteria may be restrictive given we matched food outlets by location (within a 100-meter buffer) and similar name. Depending on the research question, relaxing matching criteria may improve results (Clary and Kestens 2013). For example, a study that calculates the number of grocery stores within a 3 km buffer may benefit from a more relaxed approach than a study that measures the nearest grocery store from home.

Accessing and obtaining Yelp data are resource intensive and require advanced data management and processing decisions, including deriving food categories from open-ended text fields. However, Yelp provides timely business attributes like operating hours, delivery, and wheelchair accessibility and insights on patrons. OSM data are well documented and relatively straightforward to obtain (Ferster et al. 2019; “OpenStreetMap Wiki” 2022), but OSM does not contain the level of detail available in Yelp. Finally, accessing DMTI remains a barrier, restricted to universities and other entities that pay to use their data services. Further research is needed to determine the potential for systematic error in datasets, in particular the effects of community science users and neighbourhood characteristics (eg, socio-demographics, urban/rural) on the completeness of OSM and Yelp data.

.....

## ***ACKNOWLEDGMENTS***

The authors would like to acknowledge Linnea Soli, Sophie Cardinal, and Fiona McClave for their excellent research assistance during the initial stage of this project. CLF, JB, CF, YK, and MW were supported by the Canadian Institutes of Health Research (CIHR) for Environments and Health: Intersectoral Prevention Research.

Submitted: February 12, 2022 AEST, Accepted: May 11, 2022 AEST



This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CCBY-SA-4.0). View this license's legal deed at <https://creativecommons.org/licenses/by-sa/4.0> and legal code at <https://creativecommons.org/licenses/by-sa/4.0/legalcode> for more information.

## REFERENCES

- Bright, Jonathan, Stefano De Sabbata, Sumin Lee, Bharath Ganesh, and David K. Humphreys. 2018. "OpenStreetMap Data for Alcohol Research: Reliability Assessment and Quality Indicators." *Health & Place* 50 (March): 130–36. <https://doi.org/10.1016/j.healthplace.2018.01.009>.
- Clary, Christelle M, and Yan Kestens. 2013. "Field Validation of Secondary Data Sources: A Novel Measure of Representativity Applied to a Canadian Food Outlet Database." *International Journal of Behavioral Nutrition and Physical Activity* 10 (1): 77. <https://doi.org/10.1186/1479-5868-10-77>.
- Daepf, Madeleine IG, and Jennifer Black. 2017. "Assessing the Validity of Commercial and Municipal Food Environment Data Sets in Vancouver, Canada." *Public Health Nutrition* 20 (15): 2649–59. <https://doi.org/10.1017/s1368980017001744>.
- de Menezes, Mariana Carvalho, Vanderlei Pascoal de Matos, Maria de Fátima de Pina, Bruna Vieira de Lima Costa, Larissa Loures Mendes, Milene Cristine Pessoa, Paulo Roberto Borges de Souza-Junior, Amélia Augusta de Lima Friche, Waleska Teixeira Caiaffa, and Letícia de Oliveira Cardoso. 2020. "Web Data Mining: Validity of Data from Google Earth for Food Retail Evaluation." *Journal of Urban Health* 98 (2): 285–95. <https://doi.org/10.1007/s11524-020-00495-x>.
- Ferster, Colin, Jaimy Fischer, Kevin Manaugh, Trisalyn Nelson, and Meghan Winters. 2019. "Using OpenStreetMap to Inventory Bicycle Infrastructure: A Comparison with Open Data from Cities." *International Journal of Sustainable Transportation* 14 (1): 64–73. <https://doi.org/10.1080/15568318.2018.1519746>.
- Folch, David C., Seth E. Spielman, and Robert Manduca. 2018. "Fast Food Data: Where User-Generated Content Works and Where It Does Not." *Geographical Analysis* 50 (2): 125–40. <https://doi.org/10.1111/gean.12149>.
- Lebel, Alexandre, Madeleine I. G. Daepf, Jason P. Block, Renée Walker, Benoît Lalonde, Yan Kestens, and S. V. Subramanian. 2017. "Quantifying the Foodscape: A Systematic Review and Meta-Analysis of the Validity of Commercially Available Business Data." *PloS One* 12 (3). <https://doi.org/10.1371/journal.pone.0174417>.
- "OpenStreetMap Research." 2022. <https://wiki.openstreetmap.org/wiki/Research>.
- "OpenStreetMap Wiki." 2022. *Map Features*.
- Steinmetz-Wood, Madeleine, Kabisha Velauthapillai, Grace O'Brien, and Nancy A. Ross. 2019. "Assessing the Micro-Scale Environment Using Google Street View: The Virtual Systematic Tool for Evaluating Pedestrian Streetscapes (Virtual-STEPS)." *BMC Public Health* 19 (1): 1–11. <https://doi.org/10.1186/s12889-019-7460-3>.
- Stevenson, A.C., C. Kaufmann, R.C. Colley, L.M. Minaker, M.J. Widener, T. Burgoine, et al. 2022. "A Pan-Canadian Dataset of Neighbourhood Retail Food Environment Measures Using Statistics Canada's Business Register." *Health Reports* 33 (2).

## SUPPLEMENTARY MATERIALS

### **Supplemental Table 1. Food outlet categories in DMTI data, 2020**

Download: <https://findingspress.org/article/35619-validity-of-food-outlet-databases-from-commercial-and-community-science-datasets-in-vancouver-and-montreal/attachment/89866.html>

---

### **Supplemental Table 2. Food outlet categories in Yelp data, summer 2021**

Download: <https://findingspress.org/article/35619-validity-of-food-outlet-databases-from-commercial-and-community-science-datasets-in-vancouver-and-montreal/attachment/89867.html>

---

### **Supplemental Table 3. Food outlet categories in OSM data, summer 2021**

Download: <https://findingspress.org/article/35619-validity-of-food-outlet-databases-from-commercial-and-community-science-datasets-in-vancouver-and-montreal/attachment/89868.html>

---